Araña web

Una araña web (en inglés, web crawler) es un programa automatizado diseñado para recorrer sistemáticamente páginas web con el objetivo de recopilar información. Estas arañas son esenciales en el funcionamiento de motores de búsqueda, ya que les permiten indexar contenido y organizarlo para que sea fácilmente accesible para los usuarios.

Características de una Araña Web

1. **Automatización**: Operan sin intervención humana, siguiendo reglas y algoritmos predefinidos.
2. **Rastreo sistemático**: Se mueven entre páginas web siguiendo enlaces (hipervínculos).
3. **Recolección de datos**: Capturan contenido de las páginas visitadas, como texto, imágenes, metaetiquetas, y estructuras HTML.
4. **Velocidad**: Son capaces de rastrear millones de páginas en un corto periodo de tiempo.
5. **Adaptabilidad**: Pueden configurarse para priorizar ciertos tipos de contenido o dominios.
6. **Respeto por robots.txt**: Por lo general, siguen las reglas especificadas en el archivo robots.txt de un sitio web, que indica qué partes del sitio pueden ser rastreadas.

Proceso de Rastreo de una Araña Web

1. **Inicio del rastreo**:
    - Se comienza con una lista de URLs iniciales llamadas *semillas*.
    - Estas pueden ser proporcionadas manualmente o seleccionadas de una base de datos.
2. **Acceso y análisis**:
    - La araña accede a una página web, descarga su contenido y lo analiza.
    - Identifica los enlaces dentro de esa página para seguirlos más adelante.
3. **Almacenamiento**:
    - Los datos extraídos de cada página se almacenan en un índice, que organiza la información para ser utilizada posteriormente.
4. **Expansión del rastreo**:

- La araña sigue los enlaces identificados, ampliando su alcance a nuevas páginas.
- Este proceso continúa de forma recursiva, a menos que se encuentre un límite, como restricciones de robots.txt o políticas del sitio.

5. **Control de redundancia**:
    - Para evitar rastrear la misma página varias veces, las arañas mantienen un registro de URLs visitadas.
6. **Priorización**:
    - Las páginas más relevantes o frecuentemente actualizadas pueden ser rastreadas con mayor frecuencia.

Ejemplos de Uso de Arañas Web

1. **Motores de Búsqueda**:
    - Empresas como Google, Bing y Yahoo utilizan arañas web para recopilar y organizar la información que aparece en sus resultados de búsqueda.
2. **Comparadores de Precios**:
    - Sitios como Skyscanner o Trivago utilizan arañas web para recopilar información sobre productos, precios y disponibilidad.
3. **Monitoreo de Medios y Redes Sociales**:
    - Herramientas como Meltwater o Brandwatch utilizan arañas para seguir menciones de marcas o palabras clave en Internet.